Published article is available at:
http://applij.oxfordjournals.org/content/early/2015/11/22/applin.amv058.abstract
doi: 10.1093/applin/amv058

When *safe* means 'dangerous': a corpus investigation of risk communication in the media

> **GENDER-BEND IN RIVERS**
> CHEMICALS which block the male hormone testosterone have been found in rivers across the country. The mixture of such anti-androgens is known to interfere with sexual development and lead to low sperm count. Now, after a three-year study of 30 effluent plants in England, scientists say it is likely such 'gender-benders' are contributing to male infertility.
> *The Daily Mirror*, January 2009

Introduction

The minute presence of endocrine disrupting compounds (EDCs) and pharmaceuticals and personal care products (PPCPs) in the water supply has been the subject of considerable interest for water and environmental science researchers over the last two decades (Snyder, Vanderford et al. 2008). In both the UK and the US, strict regulations exist to ensure that that tap water is safe to drink, but no regulations currently exist that require the removal or reduction of EDCs and PPCPs. Although current scientific knowledge suggests there is no health risk (and thus no need for regulations), there is confusion due to seemingly contradictory reports, e.g. that EDCs cause sexual abnormalities in aquatic wildlife (e.g. Ramirez, Mottaleb et al. 2007). It is only recently that the UK and US media have started to report on the potential health threat of EDCs and PPCPs. Reporting was initially sporadic, but, in the US, there was significant intensification following a media campaign instigated by a series of investigative reports by the Associated Press (Donn, Mendoza et al. 2008a, 2008b; Mendoza 2008a, 2008b), leading to congress level discussions.

Amid predictions of a shift from an expert-driven regulatory process to a more participatory process influenced by public opinion (Lofstedt 2013), it is more important than ever to understand the role of the media in shaping risk perception to help generate clear, balanced and transparent information for the general public in the wake of scientific discoveries about environmental and public health risks. For EDCs and PPCPs, a key issue is that the boundary between what constitutes a health risk can quickly become blurred by speculation in the media, with potentially unfavourable consequences. For instance, during, the Measles, Mumps and Rubella (MMR) controversy in the UK, the media publicity surrounding a later discredited journal article linking the MMR vaccine and autism saw a dramatic decrease in vaccine uptake and an increase in measles outbreaks.

Although language has been identified as having a key role in the portrayal of risk and science (Slovic 2000, Kasperson, Jhaveri et al. 2001), there are relatively few studies that specifically focus on the linguistic framing of risk. We believe a formal, systematic review of the language used to represent the threat posed by EDCs and PPCPs on official (government and water industry) and media sources will inform our understanding of the framing of emerging risks that are as of yet unconfirmed by science, and the development of communication about the nature of the risks posed by contaminants.

This study uses established techniques in corpus linguistics (Sinclair 1991, Stubbs 2001a, Biber, Connor et al. 2007) to identify, quantify and compare the most prominent terms to occur in reports on contaminants, combined with a qualitative exploration of broader linguistic patterns related to the representation of risk. Our findings form the basis of a discussion on the potential of media reports to amplify risk perceptions and risk communication about contaminants.

Investigating risk communication in the media

The term *risk* has been interpreted in very different, often polarising ways in risk research literature. One key paradigm focuses on the cognitive factors that act as heuristics in the interpretation of risk, in particular the "dread risk" factor, which includes elements such as perceived lack of control and catastrophic potential, and the "unknown risk" factor, which reflects the extent a given phenomenon is understood scientifically, its observability and familiarity (e.g. Fischoff et al. 1978, Slovic 2000). Risk is also viewed from a range of sociocultural perspectives (e.g. Douglas 1966, Beck 1992) that prioritize "the social and cultural contexts in which risk is understood, lived, embodied and negotiated" (Lupton 1999: 36). Our primary interest in the current study is an understanding of how the language used to refer to contaminants and their risks as reported by science (so-called "objective risk") can provide insights into current communication practices.

The (first) representation of the complexities of risk and the science underlying risk assessments is a task often undertaken by the mass media and, as such, media reports are often seen as instrumental in the shaping of the public's perception of risk (Wahlberg and Sjoberg 2000: 38). The scientific uncertainty that surrounds health threats is seen as vulnerable to exploitation, as it provides enough latitude to reconstruct a risk as greater than it actually is (Slovic 2000). Early investigations of media content on health hazards revealed a tendency for reports to distort or exaggerate risk (e.g. Combs and Slovic 1979), but with some conflicting results (Freudenburg, Coleman et al. 1995). More recent media content analyses focusing on the role of science in risk reporting have noted polarized reactions to scientific uncertainty (Schäfer 2011), with journalists either demonstrating a reluctance to represent it (Olausson 2009), or making it the focus of controversy (Rödder and Schäfer 2010). The media has also been linked with a simplistic representation of science, a tendency that has been linked with the amplification of public concern (Barrett and Ball 2009). Reports on health hazards were found to lack an examination of cause and effect (Wahlberg and Sjoberg 2000), and a lack of specificity in terms of health outcomes and health advice (Brittle and Zint 2003).

One limitation of content based studies on risk reporting is the tendency to focus solely on media reports, despite the fact that the provision of information to the public to limit concerns is commonplace in government agencies (Timotijevic and Barnett 2006). Another lies in the lack of attention paid to the language used to frame particular threats, despite suggestions in risk research that linguistic framing plays a key role. For instance, words that have become associated with a threat are linked to a direct impact on risk perception (Kasperson, Kasperson et al. 2003: 27), potentially contributing to its stigmatization by amplifying feelings of dread (Parkin, Ragain et al. 2006). Studies in other disciplines have focused on linguistic aspects of risk. Rundblad and colleagues (2006), comparing the linguistic features of two *Lancet* articles on the links between the MMR vaccine and autism, provide evidence of the potential for media-like language to amplify perceptions of risk. There have also been studies of the metaphoric language used to represent particular health threats (Wallis and Nerlich 2005; Washer and Joffe 2006). However, surprisingly few of the many sociological investigations of environmental and public health risks make language their primary concern, a point revisited by Hamilton and colleagues' analysis of different uses of the word *risk* (2007).

Corpus linguistics and discourse analysis

While, in its primary preoccupation with linguistic phenomena, much of corpus research so far has been 'descriptive rather than explanatory' (Stubbs 2006: 28), corpus linguistic (CL) techniques are increasingly being employed by discourse analysts from a range of fields and disciplines (Baker 2006). The combination of corpus methodology and particular analytical approaches has not always been a balanced one. In some critical discourse analysis (CDA) studies employing corpus software, CL fulfils a more ancillary function, with a tendency to rely on qualitative tools such as concordancing (e.g. Magalhaes 2006).

Other investigations drawing upon a CDA framework advocate a blend of quantitative and qualitative approaches to discourse analysis (Baker, Gabrielatos et al. 2009), as do studies in the emerging field of corpus assisted discourse studies, or CADS (Partington 2004). In a CADS approach, quantitative based techniques are used to identify other examples of an already identified phenomenon, and to uncover what Partington refers to as "non-obvious meaning", i.e. patterns not identifiable from naked eye perusal (2008: 97). The interpretation of quantitative results is facilitated by a comparative analytical approach. By comparing one discourse type to another, it is possible to give observations about the (in-) frequency of a given linguistic feature the proper contextual significance (Morley 2009: 8). At the same time, concordances, which display all the instances of a particular word with its cotext, are used to gain qualitative insights. In contrast to traditional applications of CL, which discourage the researcher from reading beyond the limited cotext provided in a list of concordances, concordance lines are typically expanded, allowing a more detailed reading of stretches of text in a similar way to qualitative, text-based discourse analysis (Partington 2008). Both the quantitative and qualitative dimensions of a CADS approach, thus, address a key criticism of corpus linguistics that it tends to disregard context (Widdowson 2000).

Comparatively few corpus investigations have specifically looked at reporting on health phenomena in the media, with fewer still drawing upon quantitative and qualitative insights within a CADS framework. While Seale and colleagues use corpus techniques to identify discourses about sleep disorders in media articles (2007), the study is sociologically rather than linguistically orientated. Koteyko and colleagues (2008) draw upon corpus techniques to provide rigour in an analysis of the discourses about the MRSA "superbug". However, the main CL technique used is concordancing, with concordance data referred to as "quantitative".

Constructing a small corpus from web-based material

The current study set out to investigate and compare the prominent linguistic features evident in media reports about EDCs and PPCPs that appeared in the mass media and outreach materials (newsletters, reports, articles, etc.) posted on the websites of water industry and public health organizations. The study follows a CADS approach, firstly, in its combination of quantitative and qualitative analysis, and, secondly, in its comparative focus, i.e. we are interested in both how the reporting on contaminants is achieved linguistically, how this language differs in media and outreach reporting and what these differences say about the orientation of media and outreach organisations to reporting on contaminant risks. Rather than drawing upon a critical framework, the explanatory dimension of the analysis primarily seeks to draw upon these findings to inform our understanding of how risk communication about emerging environmental and public health threats is achieved in the media.

As we wished to include and account for any changes in reporting before and after the upsurge in US media interest following the 2008 Associated Press reports, the time period was set from January 2006 up to the point of data collection (April 2011). We also opted to establish which sites featured articles about contaminants before constructing the corpus, thereby defining what we considered to be 'media' and 'outreach'. A *googlenews* search for media content yielded 344 sites on which potentially relevant material was located. These divided into two broad categories: traditional media, which are not exclusively web-based (e.g. newspapers, magazines, broadcasters, and journals), and new media websites, which are purely web-based. As the impact of the new media on public perception is potentially significant (Gibson 2009), it was decided to include these sites in the study. However, we filtered out some sites on the basis of three main criteria: 1) the site did not employ professional journalists, 2) the site would only have niche appeal, e.g. by targeting a specific professional group, and 3) the site was defined by an explicit agenda that was directly relevant to the presence of EDCs and PPCPs, e.g. lobbying for environmentally friendly policies. Details of these criteria are given in Table 1[1]. We are thus defining 'media' as organizations whose central professional purpose is to obtain and communicate newsworthy items to a mass audience, without explicitly advocating a particular line towards EDCs and PPCPs. Outreach texts were located on the websites of water companies,

regulators and governmental public health and environmental organizations. On the same grounds as for the media, websites from organizations whose specific agenda meant advocating a particular line towards EDCs and PPCPs, such as NGOs, were excluded.

[INSERT TABLE 1 ABOUT HERE]

We searched for relevant media and outreach content on the selected sites. For many sites, the website's internal search engine was used, but for media text searches, we primarily used the LexisNexis and Highbeam databases (www.lexisnexis.co.uk; www.highbeam.com). We selected all texts that mentioned an EDC/PPCP and a reference to tap water, but also filtering texts that the average consumer would be unlikely to read, e.g. long research reports and journal articles[2]. For both media and outreach selection, any spoken data, such as transcripts of radio and TV reports, were also excluded. The filtration process thus allowed us to achieve text comparability while maintaining a representative sample for analysis[3].

Because of its focus – a finite number of texts within a limited timeframe – the corpus is necessarily small (Lee 2008), and its representativeness was largely defined by the subject matter of the texts that were chosen. Before analysis, each corpus text was 'cleaned' of any information external to article content[4]. After this process, the Media and Outreach corpus consisted of 384 media articles and 116 outreach texts, combining to reach a total word count of 343,901. For comparative purposes, the corpus was divided into four sub corpora: UK media (19,011 words), US media (255,865), UK outreach (21,075), and US outreach (47,950). The size difference between the UK and US sub corpora is explained by the greater levels of media coverage and number of water utilities in the US. Size variations between sub corpora, within certain limits[5], do not pose an obstacle to comparative analysis. A number of measures can be used to draw useful quantitative and qualitative comparisons, including normalised frequency counts, statistical tests to calculate collocational strength and the use of concordance data. These will now be discussed.

Analysing the data

Wordsmith software (Scott 1996) enables the statistical comparison of the word frequencies in the research corpus with their occurrence in a reference corpus – usually a much larger body of texts. Words revealed by such a comparison as unusually frequent are considered 'key'. The relative 'keyness' of a particular word is defined as a value based on the statistical test used, with higher values assigned to the words which are more prominent (Scott and Tribble 2006). As the uniqueness of a corpus of texts is partly defined by the topics of individual texts, key words are often content related, giving the researcher a global insight into its 'aboutness' (Scott 2001). For instance, it was naturally assumed a large proportion of key words would be related to contaminants and drinking water. Our interest was, firstly, to determine any lexical differences across the sub corpora, and, secondly, to see what other categories of key words emerged as a way of gaining entry into a large set of data.

The spelling differences in British and American English made it necessary to generate two separate reference corpora for the key word analysis. Using LexisNexis, a randomised selection of texts was downloaded for each year of the research period (2006-2011). US news was selected for the US reference corpus and UK news for the UK reference corpus. Randomised dates were chosen for each year; to reflect the larger size of the US corpus, one date was chosen per year for the UK corpus, while two dates were chosen for the US corpus. Per date, 200 articles were randomly selected. This process yielded a corpus of 360,338 words for the UK, and of 936,665 words for the US. The texts were checked for overlap with the research corpus (using the same seed terms) and were cleaned using the same procedures described above.

A key word analysis was conducted on each sub corpus using a Log Likelihood test to calculate keyness[6]. In line with the recommendation of Rayson and colleagues (2004), the critical value was set at 15.13. In order to facilitate comparative analysis, the key words were categorised into broad semantic fields. Key words that occurred in less than 5% of the total

texts were excluded from the analysis (5% corresponds to: UK Media N=38; US Media N=346; UK Outreach N=29; and US Outreach N=86). Although these categories were not pre-defined, as reports were selected on the basis of similar content, and key words tend to be content related rather than grammatical, these categories were often predictable, e.g. "types of water", "contaminants". However, for some key words, e.g. grammatical words, the word/category association could only be meaningfully defined by referring to the original context. Such categorisations were discussed by three members of the research team. As the initial categories were (deliberately) very broad, they were later divided into subcategories. So that meaningful comparisons could be made due to the different sizes of the sub corpora, the number of occurrences of each word was normalised to a count per 1,000 words (Rundblad 2007).

We also investigated phraseological and broader rhetorical patterns of key words. Firstly, we explored collocates by calculating the most frequent words to occur within a five word span, and ranking them according the strength of their relationship to the search term. Mutual Information Score (Oakes 1998) was used to calculate collocational strength, with the cut-off set at 3.0. As we were more in interested associations between lexical words, this statistic was chosen over alternatives that place more emphasis on grammatical words[7].

Additionally, a qualitative investigation was conducted using the Concord tool in *Wordsmith*, which allows all the occurrences of key words to be displayed with their cotext (normally the five words that occur immediately to the left and right). These listings of words and cotexts, or 'concordances', allow the researcher to identify syntagmatic patterns involving the search word (Sinclair 2003). In addition, *Wordsmith* permits the researcher to shift back and forth from individual lines to the original report. This process helped confirm quantitative findings and allowed us to identify broader patterns involving key words.

Results

The small size of the UK sub corpora meant that fewer words were statistically prominent enough to be considered key: 253 key words in the UK media and 276 in UK outreach. These were compared with the 500 most prominent words automatically generated by *Wordsmith* for each of the US sub corpora. The key words in all sub corpora divided into 6 broad semantic categories, which, in turn, divided into 14 sub categories (Table 2). As key words are essentially reflective of text content, the semantic categories and subcategories can be seen to reflect the key focal points of media and outreach reporting. The spotlight is clearly on the contaminants themselves (CONTAMINANTS is by far the largest category in all four sub corpora). There is also a focus on contaminant detection and causes (DISCOVERY), where they are found (WATER TYPES), their potential risks (RISK) and possible solutions (COUNTERMEASURES). In addition, there are references to a broad range of people, organisations and institutions, mostly occurring in the context of conducting operations and presenting opinions and information relating to contaminants (PARTICIPANTS).

[INSERT TABLE 2 ABOUT HERE]

An exploration of the key words in the CONTAMINANTS and RISK categories revealed differences in the representation of contaminants as a potential health threat in media and outreach reports. These will now be discussed.

The use of contaminant terms in media and outreach texts

As shown in Figure 1, the terms that could potentially be used to refer to EDCs and PPCPs can be represented as a rough taxonomy (Rosch 1978). At the superordinate level, we find general terms at a higher degree of abstraction. Basic level terms refer to specific types of compounds (*antibiotics* are a type of drug, *pesticides* are a type of chemical, etc.), and subordinate terms refer to specific compounds by name. As, in the current context, all references ultimately fall under the category of 'EDC' or 'PPCP', there is also a super-superordinate level of terms that

classify these contaminants as a unique group, such as the acronyms *EDC* and *PPCP* in outreach texts.

[INSERT FIGURE 1 ABOUT HERE]

A range of studies have shown that, in taxonomic hierarchies, basic level terms tend to be the most privileged and most frequently used (e.g. Lazareva et al. 2010). A central reason for this is that there is often a high degree of differentiation between different basic level categories but a high degree of similarity between category members (Markman and Wisniewski 1997). We therefore expected basic level terms like *antibiotics* and *cosmetics* to be the most prominent terms in our corpus. If we compare the ten most prominent terms used to refer to contaminants in media (Table 3) and outreach (Table 4), however, it is the superordinate terms, e.g. *drugs*, *chemicals, pharmaceuticals, compounds*, *contaminants*, *substances*, and *hormones* that were the most prominent. There are also differences in the type of terms used in the media versus outreach. In both UK and US media, we find a number of very prominent basic level terms – *antibiotics* and *cytotoxic* (drugs) in the UK and *pesticide*, *antibiotics,* and *pesticide* in the US. Some subordinate level terms are also very prominent in the media, specifically *atrazine* (UK and US) and *BPA* (UK). In contrast, basic level terms do not occur prominently in outreach (none of the top ten references are basic level terms), and subordinate level references are also less frequent, particularly in US outreach. There are also differences for superordinate terms: *drugs*, the most prominent term in the media, is far less prominent in US outreach and does not occur at all in UK outreach, and we only find prominent references to *hormones* in media texts. Instead, we find a range of terms at the super-superordinate level, including the acronyms *EDCs, PPCPs* and *CECs* (Contaminants of Emerging Concern) and these acronyms spelled out, e.g. (*endocrine*) *disrupting compounds/chemicals/substances*)[8].

[INSERT TABLE 3 ABOUT HERE]

[INSERT TABLE 4 ABOUT HERE]

Additionally, an analysis of the concordance lines for contaminant terms also revealed phraseological differences. Firstly, in media texts, we observed the use of negatively charged language to pre-modify superordinate level terms. For instance, for *chemicals*, one of the most prominent terms in all four sub corpora (Tables 3 and 4), phrases occur that refer to the amount or range of contaminants (*tons/tonnes of, thousands of, cocktail of, rainbow of, laundry list of, a soup of 100,000*) or their (potential) effects (*sex changing, gender-bending, cancer, cancer causing, toxic stew of*). There is also a tendency for these phrases to co-occur with *chemicals* in areas of reports that provide content framing for readers, such as the headline (Dor 2003). Secondly, while in outreach texts, contaminants are most commonly represented by super-superordinate and superordinate terms (Table 4), the prominence of basic level terms in media texts is partly explained by the occurrence of lists of contaminants, e.g. (key words are highlighted in italics):

1) In a broad study still under way, fish collected in waterways near or in Chicago; West Chester, Pa.; Orlando; Dallas; and Phoenix have tested positive for an array of *pharmaceuticals* — analgesics, *antibiotics*, *antidepressants*, antihistamines, anti-hypertension *drugs* and anti-seizure *medications*. (*Associated Press*, 11 March 2008)

In both UK and US media, over half the instances of *antibiotics,* one of the most prominent basic level terms (see Table 3), occur in such a list. In the US media corpus, some terms, e.g. (mood) *stabilisers*, (anti-) *convulsants* only occur in lists, while we see some lists copied and pasted verbatim in different articles, e.g. the cluster *antibiotics, anti-convulsants, mood stabilizers and sex hormones* occurs in 37 different reports. Louw (1993) refers to positive and negative associations as evoked by the immediate cotext as semantic prosody. Considering that

contaminant terms by themselves are likely to have negative connotations (Doria, Pidgeon et al. 2009), and that both patterns (negative language and listing) were rare in outreach texts, there is evidence of a stronger negative semantic prosody for contaminants in media texts. Further evidence of this is explored in the following section.

Evaluating contaminant risks

By comparing the key words in the RISK category, we were able to identify differences in how the evaluation of the risks posed by contaminants was presented in media and outreach texts. In terms of lexical variation, the most striking differences were found in the IMPACT sub category (Table 5).

[INSERT TABLE 5 ABOUT HERE]

Words in this sub category refer in some way to the (potential) impact of EDCs and PPCPs on humans and wildlife. Certain types of words were common to both media and outreach texts. Firstly, a range key words, e.g. *effects, impacts, impact, causing, affecting* and *cause*, simply express a causal relationship between the threat (contaminants) and the object at risk (humans or wildlife), e.g.:

2) Fish seem to have borne the brunt of the chemicals' *effects* so far. (*Philadelphia Inquirer*, February 2006)

Other terms like *harmful, toxic* and *hazardous* connotatively express this relationship:

3) DRINKING water is being contaminated with potentially *harmful* chemicals used in shampoos, shower gels and perfumes... (*Daily Mail*, December 2007)

We also find words that refer to specific adverse outcomes like hormone disruption or its potential effects, e.g.

4) Researchers also have found "*intersex*" fish – males with eggs growing in their testes. (*USA Today*, April 2011)

5) Environmental pollutants that mimic the effects of estrogen, the female hormone, may contribute to breast *cancer*. (*Wired*, April 2007)

In these cases, the object at risk is either explicit in the immediate cotext (as in Example 4) or clearly implied from the context (as in Example 5). While we find words that refer to impacts on wildlife, e.g. *disruption*, *feminisation, feminization, abnormalities, intersex,* and *interfere*, in both media and outreach texts, only in media texts do we find words that refer to specific adverse outcomes in humans (marked in bold in Table 5). A range of negative outcomes are made explicit, including fertility problems (*reproductive*, *defects, foetuses*), associations between contaminants and particular health disorders (*cancer*, *cardiovascular*) and the creation of antibiotic resistant bacteria *(resistant, resistance, bacteria, germs*). Parkin and colleagues (2006) link the occurrence of negatively charged terms in media reports, referred to as “dread words”, to inflated risk perception of water contaminants. Specific words were linked with a specific level of dread, so that terms such as *health concern* and *unknown* were associated with moderate concern, and *miscarriage, toxic, death* and *cancer* with serious concern. Due to their largely negative associations, we would argue that most IMPACT key words are likely to fall into the moderate or serious concern categories. Also, as it is typical to find a contaminant term within a short span (see Examples 2, 3 and 5), these negative references contribute to a negative semantic prosody for contaminant terms. The prominence of key words referring to specific human impacts indicates that such associations are a more prevalent feature of media texts.

We also identified words that were central in communicating how contaminant risks are evaluated by the water industry and scientific community by exploring collocates, concordance lines and the wider context of key words. As this is typically a labour-intensive process, it was important to exercise selectivity. In choosing words for more detailed analysis, we were careful to consider their semantic category (e.g. the UNCERTAINTY and OPINION subcategories were an obvious place to start), the prominence of the words across discourse types, as well as insights based on our knowledge of risk and water research relating to contaminants[9]. This process revealed an important role for the key words *risk, concern*, *evidence,* and *safe*.

Table 6 shows the ten strongest collocates for *risk* and *concern* in US media and US outreach, calculated using a minimum frequency of 5 occurrences[10]. We set up the software to identify the words most strongly associated with the node word based on the occurrence of words within five word span to its left and right. Among the most prominent collocates, we find several that mark the level of risk or concern as more or less intense. Negative certainty markers (*not* or *no* + *concern*) delineate a lack of *risk/concern*:

6) Detections, therefore, do **not** necessarily indicate a *concern* to human health…(US outreach sub corpus: *US Geological Survey*)

There are also words that hedge the amount of risk (Example 7), words that mark *risk* as something that is intensifying (Example 8) or apparent (Example 9), and *concern* as somehow validated (Example 10) (bold and italics our emphasis throughout):

7) …the levels are so small that they pose **little** *risk*. (*Fox News*, February 2010)

8) …such drug concoctions can **heighten** the *risk* of cancer in humans. (*Boston Globe*, September 2008)

9) For several decades, federal environmental officials and nonprofit watchdog environmental groups have focused on regulated contaminants – pesticides, lead, PCBs – which are present in higher concentrations and **clearly** pose a health *risk*. (*Katu.com*, March 2008)

10) …there is **genuine** *concern* that these compounds… …could be causing impacts to human health or to aquatic organisms. (*Foodconsumer*, January 2011)

There is a difference in the type of collocates to occur in the two sub corpora. In US outreach, negative markers are the most prominent device used in risk evaluations involving the words *risk* and *concern* (Table 6), and intensifying terms (e.g. *heighten, clearly, genuine*) do not occur. In US media, while intensifying terms are prominent (*heighten/clearly* + *risk* and *genuine/growing/increasing* + *concern*), we also find negative markers (*not* + *risk*) and words that mark *risk* as less or more probable (*little* vs. *potential*+ *risk*). These diverse and even conflicting representations of risk in the US media contrast with the consistent focus on the lack or absence of risk in US outreach texts.

[INSERT TABLE 6 ABOUT HERE]

The conflicting risk evaluations in US media can be linked to the practice of explicitly contextualising risk evaluations as the opinions of different authorities. Amongst the strongest collocates of *risk* and *concern* in the US media are words that refer to organisations or people – the pharmaceutical manufacturer *GlaxoSmithKline* (+ *risk*) and the former Water Administrator at the Environmental Protection Agency Peter Silva (*Peter* and *Silva* + *concern*). Also highly ranked are *raises* (+ *risk*) and *voicing* and *recognise* (+ *concern*), which are all used in the context of associating a particular authority with a particular assessment:

11) …said Conrad Volz, a University of Pittsburgh scientist whose research **raises** questions about the *risk* of eating fish from waters contaminated with sex hormones (*Seattle Times*¸ December 2009)

12) …despite EPA director Lisa Jackson publicly **voicing** *concern* about the chemical. (*Associated Press*, March 2010)

13) "We **recognize** it is a growing *concern* and we're taking it very seriously," said Benjamin H. Grumbles… (*Washington Post*, March 2008)

In contrast, the risk assessing authority in US Outreach is barely present on the surface of the text, reflected in the high ranking of grammatical words (*to, there, not, no, of*) in the list of collocates. For instance, *there* and *risk* occur in agentless statements about the level of risk, e.g.

14) …the best research to date does not demonstrate that **there** is a human health *risk.* (US outreach sub corpus: *WSSC Water*)

In both UK and US outreach texts, we also find *to* (+ *risk*) and *of* (+ *concern*) in nominalised phrases that conflate the risk assessor and their evaluation with either the object at risk (*risk to human health*) or the threat itself (*contaminants of concern* and *constituents of emerging concern*).

The important role for science in the representation of risk in media texts is underlined by the fact that *evidence* has keyness in the US media. Concordances reveal that *evidence* typically occurs in claims about the (non-) existence or discovery of evidence that contaminants are harmful, e.g.

15) …some even **finding** *evidence* of reduced sperm count in men from agricultural regions of the U.S. (*Scientific American*, March 2010)

16) So far, there is **no** *evidence* that tap water from the Potomac is unsafe to drink, according to Jacobus and officials at other area utilities. (*CBS News*, January 2010)

The use of *evidence* in these claims is tantamount to a risk evaluation, whereby a lack of evidence suggests a lack of risk and vice versa. 63 out of 129 occurrences of *evidence* are in claims about the existence/discovery of evidence compared with 43 in claims about a lack of evidence. Broadening the concordance lines revealed a tendency for the "no evidence" claims to be followed by a counterclaim, e.g.

17) There is **no** solid *evidence* yet that trace amounts of anticholesterols, mood stabilizers, hormones, and other substances are bad for human health. **But there are worrisome signs.** (*Boston Globe*, March 2008).

In each counterclaim sequence, the first claim is undermined in some way by the second claim. This contrast is accentuated by the use of adversative discourse markers, like *but* in Example 18, as well as other conjunctives (*however, although, while*) and adverbs (*still, yet*). The counterclaim sequence can therefore be summarised as:

(*although*, *while*) CLAIM 1 + (*but*, *however*, *still*, *yet*) CLAIM 2

Counterclaim patterns were also identified for other key words in the RISK category. For instance, the word *safe* was found to occur in claims stating that either a particular contaminant or tap water containing contaminants is safe, e.g.

…utilities insist their water is *safe*.

Expanding the concordance lines revealed the same pattern as for *evidence,* e.g.

18) …utilities insist their water is *safe*. **But the presence of so many prescription drugs… …in so much of our drinking water is heightening worries among scientists of long-term consequences to human health.** (*Associated Press*, 10 March 2008)

The overwhelming tendency is for counterclaims to follow claims that evaluate contaminant risk as less intense. Approximately 70% (29/43) of "no evidence" claims, and just over 85% (92/106) of "water is safe" claims are followed by a counterclaim. Although counterclaiming was a less observed phenomenon for risk evaluations in outreach texts, when it does occur, in a direct reversal of the media pattern, the more attenuating claim comes second:

19) There is *evidence* to show that EDCs can pose a hazard to some wildlife and marine studies by Defra (EDMAR) have identified this risk may be significant for wildlife under drought conditions. **However these environmental concerns are quite separate from, and should not be related to drinking water.** (UK outreach sub corpus: *Drinking Water Inspectorate*)

Counterclaiming, thus, operates as a rhetorical device that draws together contrasting evaluations of risk, typically from different authorities, with the first claim acting as a "dissenting" view that is overridden by the second. In media texts, where this is pattern is more common, the dissenting slot tends to be occupied by the more attenuating evaluation, even though such evaluations more accurately reflect the general scientific consensus at the time.

Conclusion

Our study set out to compare the language used to report the discovery of contaminants and their potential risks in media and outreach reporting. The findings provide linguistic evidence to support the range of content-based studies pointing toward the media's role in the social amplification of risk (Brittle and Zint 2003; Olausson 2009; Rödder and Schäfer 2010; Schäfer 2011). Although reporting on the same phenomenon, the language used to refer to contaminants and their risks differs considerably in the two discourse types, in terms of lexical choice, semantic prosody and how the risks are represented rhetorically.

The more persistent syntagmatic associations of the negative terms occurring in the media result in a stronger negative semantic prosody for contaminants. This has two implications for risk communication. Firstly, the negatively charged language may lead to the stigmatization of contaminants (Kasperson 2003), potentially influencing the public's affective response (Loewenstein 2001, Setbon, Raude et al. 2005). Secondly, the more specific and frequent references to severe human health outcomes (cancer, fertility problems, etc.) mean that media reports may be more likely to have a negative impact on risk perception.

The general scientific consensus in the field of water research at the time is that there is, in fact, no human health risk (Khiari 2007). Because of the way scientific claims are presented, i.e. with the mainstream view represented as the "dissenting voice", readers of media reports may be persuaded to interpret the lack of evidence for a health threat as simply preliminary to its discovery. It might be tempting to cite the manipulation of language and source material in media texts as evidence of a journalistic imperative "to sell a story", though, the reality, for which there is little space to discuss here, is likely to be far more complex[11]. Authorial motivations aside, the growing influence of public opinion in regulatory processes (Lofstedt 2011) means the potential cognitive and affective impact of media reports has strong implications for how regulators and water companies respond to the presence of contaminants. Although an objective representation of risk is clearly an impossibility, this case demonstrates a

potential relationship between the social amplification of risk and the linguistic dramatisation of what is, at best, insubstantial evidence about a possible health threat.

Although, linguistically, outreach reports present a less inflammatory account of contaminant risks, they are likely to be less accessible to information seekers. Searches for outreach advice about emerging health risks are typically prompted by media publicity (Kasperson et al. 2003). Due to the differences in contaminant terminology between media and outreach texts, people entering the prominent media terms on internet search engines are more likely to encounter other media reports. Considering that the most popular internet search engine queries consist of single words that are non-technical and in common usage (Spink, Wolfram et al. 2001), searches are more likely to use basic or subordinate level words such as *antibiotics* and *atrazine*, than super-superordinate terms, such as *EDC* or *PPCP*, which are mainly found in outreach reports.

The prominent references to scientific evidence and opinion in both media and outreach reporting highlight the importance of reporting on research findings in a balanced and responsible manner. There are key challenges in communicating whether particular claims are based solely on (individual) observations of statistical significance, or whether this significance is adjudged to be important enough to justify immediate action in the interest of public health. This study has shown that there is an important role for linguistic and rhetorical structures in distinguishing between the two scenarios. Integrating linguistic analysis in risk communication research may, therefore, facilitate strategies for the design of outreach materials that offer a more accessible, balanced and non-inflammatory representation of the science and risk behind emerging threats to public health, such as EDCs and PPCPs.

The above observations demonstrate the applicability of corpus linguistic techniques for discourse analysis and risk communication research. As we cannot assume that individual articles will be read carefully from start to finish (O'Halloran 2003), uncovering typical and frequent occurrences of words as opposed to what happens to occur (Stubbs 2001b), reveals the type of language readers are most likely to be exposed to. However, as corpus analysis relies on tests of statistical significance, the much smaller size of the UK sub corpora limited the comparison between the language in UK and US reporting. For instance, similar patterns of claim and counter claim were identified in UK media texts based on a qualitative exploration of whole texts, but these patterns were not as clearly evident as in the US media simply based on an analysis of the collocates and concordances of key words. Also, it must be noted that the study's inclusion of new media did not extend to reports that appeared on the websites of potentially influential health and environmental groups, and spoken media was also not explored. Whether these other types of media will reveal similar patterns of language might be the subject of further empirical investigation.

Furthermore, while this study has identified a range of language in the media that may be directly linked to the social amplification of risk, it remains to be seen if individual features will have the predicted effect on those reading reports. Further research might involve the use of cognitive linguistic and psycholinguistic tools to identify the associations people make when presented with the most prominent terms, as well as an assessment of risk perception based on reactions to the rhetorical and linguistic features found.

Funding

This paper forms part of the "Consumer Perceptions and Attitudes towards EDCs and PPCPs in Drinking Water" project, funded by the Water Research Foundation (#4323). This is a 501(c)3 nonprofit organization, with a research agenda governed by a Board of Trustees and a Public Council that includes well renowned consumer organisations and advocacy groups.

Acknowledgements

We particularly wish to thank Chris Tribble, the editor and the anonymous reviewers for reading and commenting on this paper. Additional thanks are due to Lisa Ragain and Jennifer

Breedlove for advice on US sources, to the Water Research Foundation Project Advisory Committee, Linda Reekie for continued support throughout the project and to Roseanna Cooke for cleaning the corpora.

---

[1]This information was ascertained by qualitative exploration of each individual website, e.g. by clicking on the *about* tab.

[2]Assessing text suitability was based on a qualitative evaluation of genre and register, a procedure that benefited from the insights of a member of the research team who had extensive experience working in US water industry communications, and of our partners in the UK water industry.

[3]A more detailed explanation of these processes can be found in Rundblad and colleagues 2013.

[4]For instance, LexisNexis adds tags before and after texts that give information about the publication, the author and the date, and copying and pasting html texts can leave unwanted characters and image captions.

[5]Paul Rayson, developer of Wmatrix, recommends corpora sizes should not be out by more than a factor of 10 (O'Halloran 2011: 177).

[6]Log Likelihood tests have been found to reveal accurate results even with smaller samples (Rayson et al. 2004).

[7] For a useful discussion of different statistical measures used to calculate collocational strength see Baker 2006: 95f.

[8] Here and throughout, brackets indicate words that occur with the key word in 90% or more cases.

[9] We would argue that, even in "corpus-driven" studies (Tognini-Bonelli 2001), intuitions based on prior knowledge can be an asset to the researcher, as long as intuitive judgements are grounded in the quantitative findings and affirmed by qualitative exploration of concordance lines.

[10]The smaller size of the UK sub corpora did not provide enough instances of *risk* or *concern* for meaningful comparison.

[11]For example, see Klinenberg's account (2003) of the media's handling of the 1995 Chicago heatwave.

References

Baker, **P**. 2006. *Using corpora in discourse analysis*. Bloomsbury Publishing.

Baker, **P**., **Gabrielatos**, **C**., **Khosravinik**, **M**., **Krzyżanowski**, **M.**, **McEnery**, **T**., and **Wodak**, **R.** 2008. 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press'. *Discourse and Society 19*/3: 273-306.

Barrett, **M.** and **D. Ball**. 2009. *Experts and Public Risk*, Risk and Regulation Advisory Council.

Beck, **U.** 1992. *Risk society: Towards a new modernity* 17. Sage.

Biber, **D.**, **U. Connor** and **T. Upton**. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. John Benjamins.

Brittle, **C.** and **M. Zint**. 2003. 'Do newspapers lead with lead? A content analysis of how lead health risks to children are covered.' *Journal of Environmental health* 65: 17-22.

Combs, **B.** and **P. Slovic**.1979. 'Newspaper coverage of causes of death.' *Journalism Quarterly* 56.

Donn, **J.**, **M. Mendoza** and **J. Pritchard**. 2008a. 'No standards to test for drugs in water.' *Associated Press* Mar 11.

Donn, **J.**, **M. Mendoza** and **J. Pritchard.** 2008b. 'Pharmaceuticals found in drinking water of 24 major metro areas, 34 say no testing.' *Associated Press* Mar 17.

Doria, **M. F.**, **N. Pidgeon** and **P. R. Hunter**. 2009. 'Perceptions of drinking water quality and risk and its effect on behaviour: A cross-national study.' *Science of the Total Environment* 407: 5455-5464.

Douglas, **M.** 2003. *Purity and danger: An analysis of concepts of pollution and taboo*. Routledge.

Dor, D. 2003. 'On newspaper headlines as relevance optimizers.' *Journal of Pragmatics* 35/5: 695-721.

Fischhoff, **B.**, **Slovic**, **P.**, **Lichtenstein**, **S.**, **Read**, **S.**, and **Combs**, **B.** 1978. 'How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits'. *Policy sciences* 9/2: 127-152.

Freudenburg, **W. R., C. Coleman**, **J. Gonazales** and **C. Helgeland**. 1995. 'Media Coverage of Hazards: Analyzing the Assumptions.' *Risk Analysis* 16/1: 31-41.

Gibson, **R. K.** 2009. 'New media and the revitalisation of politics.' *Representation* 45/3: 289-299.

Hamilton, **C.**, **S. Adolphs** and **B. Nerlich**. 2007. 'The meanings of 'risk': a view from corpus linguistics.' *Discourse and Society* 18/2: 163-181.

Kasperson, **J. X.**, **R. E. Kasperson**, **N. Pidgeon** and **P. Slovic**. 2003. 'The social amplification of risk: Assessing fifteen years of research and theory' in N. Pidgeon, R. E. Kasperson and P. Slovic (eds.): *The social amplification of risk*. Cambridge, Cambridge University Press: 13-46.

Kasperson, **R.**, **N. Jhaveri** and **J. X. Kasperson**. 2001. 'Stigma and the Social Amplification of risk. 2001 : Toward a framework of analysis' in J. Flynn, P. Slovic and H. Kunreuther (eds.): *Risk, media and stigma*. London, Earthscan: 9-27.

Khiari, **D**. 2007. 'Endocrine disruptors, pharmaceuticals and personal care products in drinking water: an overview of AWWARF research to date.' *Drinking Water Research* 17/2: 2-7.

Klinenberg, **E.** 2003. *Heat wave: A social autopsy of disaster in Chicago*. University of Chicago Press.

Koteyko, **N.**, **B. Brown** and **P. Crawford**. 2008. 'The dead parrot and the dying swan: The role of metaphor scenarios in UK press coverage of avian flu in the UK in 2005–2006.' *Metaphor and Symbol* 23: 242-261.

Lazareva, **O. F.**, **Soto**, **F. A.**, and **Wasserman**, **E. A.** 2010. 'Effect of between-category similarity on basic level superiority in pigeons'. *Behavioural processes* 85/3: 236-245.

Lee, **D**. 2008. 'Corpora and discourse Analysis: New Ways of Doing Old Things' in V. Bhatia, J. Flowerdew and H. Jones (eds.): *Advances in Discourse Studies*. London, Routledge: 86-99.

Lofstedt, **R.** 2013. 'Communicating Food Risks in an Era of Growing Public Distrust: Three Case Studies'. *Risk Analysis* 33: 192-202.

Loewenstein, **G. F. W.**, **E. U. Weber**, **C. K. Hsee**, and **N. Welch**. 2001. 'Risk as feelings.' *Psychological Bulletin* 127/2: 267-286.

Louw, **B.** 1993. 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies' in M. Baker, G. Francis and E. Tognini-Bonelli (eds.): *Text and Technology: In Honour of John Sinclair*. Philadephia/Amsterdam, John Benjamins: 157-176.

Lupton, **D.** 1999. *Risk*. Routledge.

Magalhaes, **C.M.** 2006 'A Critical Discourse Analysis Approach to News Discourses and Social Practices on Race in Brazil'. *DELTA* 22/2: 275–301.

Markman, **A. B.**, and **Wisniewski**, **E. J.** 1997. 'Similar and different: The differentiation of basic-level categories'. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23/1: 54.

Mendoza, **M.** 2008a. 'AP water probe prompts Senate hearings.' *Associated Press* Mar 12.

Mendoza, **M.** 2008b. 'Providers, researchers keeping secrets.' *Associated Press* Mar 10.

Morley, **J.** 2009. 'Introduction: A description of CorDis' in J. Morley and P. Bayley (eds.): *Corpora and discourse*. Bern, Peter Lang: 11-20.

O'Halloran, **K. A.** 2003. *Critical discourse analysis and language cognition*. Edinburgh, Edinburgh University Press.

O'Halloran, K. A. 2011. 'Investigating argumentation in reading groups: Combining manual qualitative coding and automated corpus analysis tools.' *Applied Linguistics* 32/2: 172-196.

Oakes, **M.** 1998. *Statistics for corpus linguistics*. Edinburgh, Edinburgh University Press.

Olausson, **U.** 2009. 'Global Warming – Global Responsibility? Media Frames of Collective Action and Scientific Certainty.' *Public Understanding of Science* 18: 421-436.

Parkin, **R.**, **L. Ragain**, **R. Bruhl**, **H. Deutsch** and **P. Wilborne-Davis**. 2006. *Advancing Collaborations for Water-Related Health Risk Communication*. Denver, CO., American Water Works Research Foundation.

Partington, **A.** 2004. 'Corpora and discourse, a most congruous beast' in A. Partington, J. Morley and L. Haarman (eds.). *Corpora and Discourse*. Bern, Peter Lang: 11–20.

Partington, **A.** 2008. 'The armchair and the machine: corpus-assisted discourse research'. *Corpora for university language teachers* 74. Peter Lang: 95.

Ramirez, **A. J.**, **M. A. Mottaleb**, **B. W. Brooks** and **C. K. Chambliss**. 2007. 'Analysis of Pharmaceuticals in Fish Using Liquid Chromatography-Tandem Mass Spectrometry.' *Analytical Chemistry* 79/8: 3155-3163.

Rayson, P., D. Berridge, D. and **B. Francis**, 2004. 'Extending the Cochran rule for the comparison of word frequencies between corpora'. *JADT* 7: 1-12.

Rödder, **S.** and **M. S. Schäfer**. 2010. 'Repercussion and resistance: An Empirical Study in the interrelation between science and mass media.' *Communications* 35: 249-267.

Rosch, **E.** 1978. 'Principles of categorization' in E. Rosch and B. B. Lloyd (eds.): *Cognition and categorization*. Hillsdale: Lawrence Erbaum Associates: 27-48.

Rundblad, **G** 2007. 'Impersonal, General, and Social: The Use of Metonymy Versus Passive Voice in Medical Discourse.' *Written Communication* 24/3: 250-277.

Rundblad, **G**, **P. A. Chilton** and **P. R. Hunter**. 2006. 'An Enquiry into Scientific and Media Discourse in the MMR Controversy: Authority and Factuality.' *Communication and Medicine* 3/1.

Rundblad, **G**, **C. Tang**, **O. Knapton**, **L. Ragain**, **M. Myzer**, **A. Tytus, J. Breedlove** and **R. Cooke**. 2013. *Consumer Petceptions and Attitudes Toward EDCs and PPCPs in Drinking Water*. Water Research Foundation.

Schäfer, **M. S.** 2011. 'Sources, characteristics and effects of mass media communication on science. A review of the literature, current trends and areas for future research.' *Sociology Compass* 5/6: 399-412.

Scott, **M.** 1996. *Wordsmith*. Oxford, Oxford University Press.

Scott, **M.** 2001. 'Comparing corpora and identifying keywords, collections, frequency distributions through the wordsmith tools suite of computer programs' in M. Ghadessy, A. Henry and R. L. Roseberry (eds.): *Small Corpus studies and ELT: theory and practice*. Amsterdam, John Benjamins: 47-70.

Scott, **M.** and **C. Tribble**. 2006. *Textual Patterns. Key words and corpus analysis in language education*. Amsterdam/Philadelphia, John Benjamins.

Seale, **C.**, **S. Boden**, **S. Williams**, **P. Lowe** and **D. Steinberg.** 2007. 'Media constructions of sleep and sleep disorders: A study of UK national newspapers.' *Social Science and Medicine* 65: 418-430.

Setbon, **M., J. Raude**, **C. Fischler** and **A. Flahault**. 2005. 'Risk Perception of the 'Mad Cow Disease' in France: Determinants and Consequences.' *Risk Analysis* 25/4: 813-826.

Sinclair, **J.** 1991. *Corpus, Concordance, Collocation*. Oxford, Oxford University Press.

Sinclair, **J.** 2003. *Reading Concordances*. Harlow, Pearson Longman.

Slovic, **P. E.** 2000. *The Perception of Risk*. London, Eartchscan Publications.

Snyder, **S.**, **B. J. Vanderford**, **J. Drewes**, **E. Dickenson**, **E. M. Snyder**, **G. M. Bruce** and **R. C. Pleus.** 2008. *State of Knowledge of Endocrine Disruptors and Pharmaceuticals in Drinking Water*. Denver, CO: AWWA Research Foundation.

Spink, **A.**, **D. Wolfram**, **M. B. J. Jansen** and **T. Saracevic**. 2001. 'Searching the web: The public and their queries.' *Journal of the American Society for Information Science and Technology* 52/3: 226-234.

Stubbs, **M.** 2001a. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford, Blackwell.

Stubbs, **M.** 2001b. 'Texts, Corpora, and Problems of Interpretation: A Response to Widdowson.' *Applied Linguistics* 22/2: 149-172

Stubbs, **M.** 2006. 'Corpus analysis: the state of the art and three types of unanswered questions' in G. Thompson and S. Hunston (eds.): *System and Corpus*. London, Equinox: 15-36.

Timotijevic, **L.** and **D. J. Barnett** 2006. 'Managing the possible health risks of mobile telecommunications: Public understandings of precautionary action and advice'. *Health, risk and society* 8: 143-164.

Tognini-Bonelli, **E.** 2001. *Corpus Linguistics at Work*. Amsterdam, Benjamins.

Wahlberg, **A.** and **L. Sjoberg**. 2000. 'Risk perception and the media.' *Journal of Risk Research* 3/1: 31-50.

Wallis, **P.**, and **Nerlich**, **B.** 2005. 'Disease metaphors in new epidemics: the UK media framing of the 2003 SARS epidemic'. *Social science and medicine* 60/11: 2629-2639.

Washer, **P.**, and **Joffe**, **H.** 2006. 'The "hospital superbug": social representations of MRSA'. *Social Science and Medicine*. 63/8: 2141-2152.

Widdowson, **H.G.** 2000. 'On the Limitations of Linguistics Applied'. *Applied Linguistics* 21/1: 3–25.

Figure 1: Taxonomy of contaminant terms (adapted from Rundblad et al. 2013)

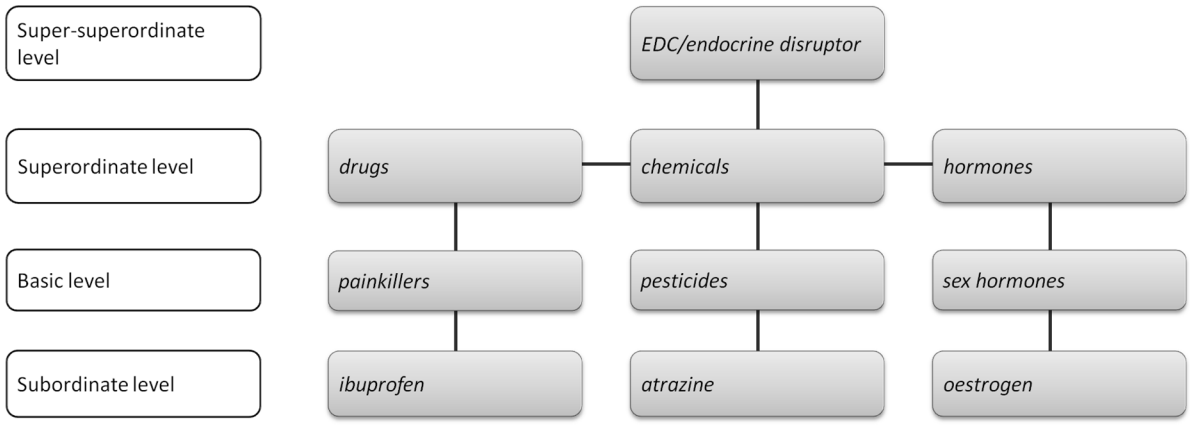

Table 1: Inclusion and exclusion criteria used to define 'media'
(adapted from Rundblad et al. 2013)

Type of journalism	We included sites providing or sourcing original articles written by journalists regularly employed by the site and excluded sites that primarily relied on participatory journalism or expert blogs. We thus filtered out academic journal articles, health blogs by doctors and letters to the editor from the general public whilst retaining blogs by journalists on newspaper websites.
Advocacy Bias	We excluded websites that were defined by specific, explicit agenda that predisposed writers to advocate a position in relation to EDCs and PPCPs as water contaminants, for example, websites belonging to environmental groups or websites advertising water filters. Websites that were included were those that represented a middle ground, reporting on a particular subject area with popular appeal, e.g. news, science, health.
Niche Appeal	Websites aimed at small, highly-specific interest or professional groups, e.g. dog owners, lawyers, were excluded. Websites reporting on topics with popular, widespread appeal were included, e.g. those focusing on health issues, current affairs or science.

Table 2: Key word categories

Category	**Sub categories**	**Examples from US media sub corpus**
CONTAMINANTS	WATER CONTAMINANTS	*pharmaceuticals, contaminants, antibiotics*
	QUANTITY	*concentrations, levels, traces, amounts*
DISCOVERY	CONTAMINATION CHANNELS	*sewage, wastewater, flushing, toilets*
	RESEARCH	*found, detected, research, samples*
WATER TYPES	WATER SUPPLY	*water*, *drinking* (water), *tap* (water)
	WATER SOURCES	(the) *environment, rivers, lakes*
PARTICIPANTS	AUTHORITIES	*EPA, studies, scientists, Grumbles*
	OPINION	*concerns , concern, says*
RISK	UNCERTAINTY	*risk, potential, evidence*
	EXPOSURE	*contaminated*, *exposure*, *exposed*

	IMPACT	*effects, human, wildlife*
COUNTERMEASURES	DISPOSAL	*collection, sites, (kitty) litter*
	TREATMENT	*treatment, treated, process,*
	REGULATIONS	*exceeded, regulations, standards*

Table 3: Ten most prominent key words referring to contaminants in the UK and US media

UK			US		
Term	Keyness	Norm/n	Term	Keyness	Norm/n
drugs	581.35	6.7/127	*drugs*	3933.04	5.8/1482
chemicals	453.92	4.4/83	*pharmaceuticals*	3827.39	4.9/1255
pharmaceuticals	281.49	2.5/47	*atrazine*	1857.36	2.4/603
BPA	203.61	1.8/34	*chemicals*	1744.04	2.4/623
atrazine	179.65	1.5/30	*pesticides*	879.22	1.1/293
cytotoxic (drugs)	173.66	1.5/29	*compounds*	850.79	1.1/292
compounds	137.72	0.4/23	*contaminants*	834.45	1.1/271
contaminants	137.72	0.2/23	*hormones*	775.64	1.0/256
antibiotics	95.8	0.8/16	*antibiotics*	614.84	0.8/212
hormones	65.86	0.6/11	*pesticide*	511.09	0.6/166

Table 4: Ten most prominent key words referring to contaminants in UK and US outreach (adapted from Rundblad et al. 2013)

*(*compounds*/*chemicals*/*substances*)

UK			US		
Term	Keyness	Norm/n	Term	Keyness	Norm/n
chemicals	432.08	3.9/82	*pharmaceuticals*	2065.48	7.2/347
substances	399.83	3.3/69	*compounds*	1217.15	4.4/209
pharmaceuticals	330.26	2.7/57	*contaminants*	1173.32	4.0/194
compounds	243.32	2.0/42	*PPCPs*	1022.03	3.5/169
EDCs	225.93	1.1/39	*chemicals*	758.63	3.0/146
(endocrine) *disrupters*	144.82	1.2/25	*emerging* (contaminants)	538.21	2.3/110
oestrogens	139.02	1.1/24	*EDCs*	519.95	1.8/86
EDC	133.23	1.1/23	*drugs*	450.95	2.5/120
(endocrine) *disrupting**	119.32	1.0/22	*medications*	436.58	1.8/86
oestradiol	110.06	0.9/19	*CECs*	411.10	1.4/68

Table 5: Key words referring to contaminant effects ranked in order of keyness[a]
(adapted from Rundblad et al. 2013)

[a]words in bold refer to specific contaminant effects on humans [b](n/normalised count)

UK media	*effects* (51/2.7)[b], ***cancer*** (28/1.5), *impacts* (17/0.9), ***bacteria***,(antibiotic/drug) (15/0.8), ***resistant*** (15/0.8), *harmful* (11/0.6), *feminisation* (8/0.4), *toxic* (13/0.7), ***foetuses*** (5/0.3), ***resistance*** (9/0.5), ***cells*** (12/0.6), ***cardiovascular*** (disease) (6/0.3), *disruption* (9/0.5), *toxicity* (6/0.3)
US media	*effects* (363/1.4), *toxic* (78/0.3), *intersex* (66/0.3), *harm* (114/0.4), ***reproductive*** (82/0.3), (birth) ***defects*** (77/0.3), *harmful* (58/0.2), *impacts* (60/0.2), *feminized* (39/0.2), ***bacteria*** (56/0.2), ***cancer*** (107/0.4), *hazardous* (55/0.2), ***germs*** (34/0.1), *abnormalities* (22/0.1), *causing* (56/0.2), *eggs* (70/0.3), *sperm* (39/0.2), *organs* (39/0.2), *testes* (24/0.1), *interfere* (27/0.1), *kidney* (27/0.1), *affected* (63/0.2), *impact* (94/0.4), ***resistance*** (38/0.1), ***resistant*** (29/0.1), *affecting* (39/0.2)
UK outreach	*effects* (55/2.6), *feminisation* (14/0.7), *reproductive* (9/0.4), *harmful* (8/0.4), *disruption* (7/0.3), *gonadal* (7/0.3)
US outreach	*impacts* (34/0.7), *disruption* (13/0.3), *intersex* (16/0.3), *harmful* (15/0.3), *reproductive* (11/0.2), *hazardous* (18/0.4), *kills* (11/0.2), *cause* (25/0.5), *toxicity* (7/0.1), *impact* (26/0.5)

Table 6: Ten strongest collocates of *risk* and *concern* in US media and US outreach ranked according to Mutual Information score

	collocates of *risk*				collocates of *concern*			
	US media		US outreach		US media		US outreach	
1	*heighten*	10.5	*due*	9.3	*voicing*	10.6	*constituents*	9.2
2	*concoctions*	10.5	*assessment*	9.2	*genuine*	10.6	*emerging*	7.3
3	*glaxosmithkline*	9.9	*extremely*	7.6	*recognize*	9.8	*no*	6.3
4	*clearly*	9.6	*there*	7.1	*peter*	9.7	*human*	7.6
5	*assessment*	8.7	*low*	6.6	*silva*	9.6	*contaminants*	5.7
6	*eating*	8.6	*potential*	6.5	*share*	8.9	*health*	5.0
7	*pose*	8.5	*health*	6.5	*growing*	8.7	*of*	4.7
8	*raises*	8.4	*human*	6.4	*meanwhile*	8.7	*levels*	4.6
9	*contribute*	8.1	*to*	4.7	*chronic*	8.2	*not*	4.5
10	*little*	7.8	*is*	4.7	*increasing*	7.9	*environment*	4.5